基于汉语声学模型的维语语音识别系统

1 背景介绍

目前，语音识别市场由国外公司和机构占据了很大份额，系统以英语为主，我国紧跟语音识别领域的最新研究成果并基本与之保持同步。汉语语音技术的广泛应用，使大家看到语音技术的广大市场前景。

在新疆，维吾尔族是自治区的自治民族，少数民族尤其是维吾尔族在新疆人口中占有很大的比例。新疆地区官方语音是汉语和维语，由于维吾尔语语音特性，维、汉语之间的语言差异很大，正是这种少数民族的构成、人口与语言文字状况，使少数民族语言文字信息技术的开发与应用成为新疆信息化建设当中不可或缺的一个重要方面，也是国家信息化的基础之一。而研究维吾尔语的语音识别系统是新疆信息化建设的内容之一。具有重大的研究意义。到目前为止，国外无一机构（包括微软、IBM 等跨国公司）从事维语信息处理及维语语音识别系统的开发，所以目前国际上在此领域的研发也是一片空白。

同时，新疆地区的哈萨克民族和新疆周边的中亚国家，他们的语言文字和维吾尔语十分相似，维吾尔语的语音识别技术不仅在新疆有很广的应用前景，也能为这些语言的相关研究提供技术参考。语音识别系统广泛的应用市场和维吾尔族用户所占比例表明维吾尔文语音识别系统研究开发工作的必要性，其市场也是不容忽视的。

2 问题描述

语音识别系统本身具有快捷、方便、智能、适用设备和人群广泛等特点，而维语语音识别系统恰能弥补国内外在维语信息处理这一块的技术空白。不仅能给新疆地区语音研究领域建立坚固的基础，还能给少数民族用户带去高新技术智能产品。然而，研发与汉语发音差异巨大的维语语音识别系统带来了一些问题。首先，对于发音变化的比较好的解决方法是针对不同的语言训练不同的声学模型，但是，由于维语属于少数民族语言，数据的采集不易，导致训练数据有限，不能完全覆盖维语的特点，这会引发训练出的声学模型与实际发音不匹配的问题；第二，国内外解决目标语言训练样本短缺的方法虽然繁多，但是要么需要经过较长时间的训练，要么训练不充分，都没法从根本上提升系统效率，没有一种统一的、高效的方法。最后，在效果较好的基于原有大词汇量连续语音识别系统声学模型基础上，训练新语音识别系统的方法里面，没有一个合适的评价方法指出构造的模型层次结构，模型层次结构构造具有随机性。这些环环相扣的问题都给具有稀疏训练数据的维语语音识别系统开发带来了阻碍。

本发明提出一种基于汉语声学模型的维语语音识别系统，在基于汉语的基础上解决维语语料不足的问题，并且利用原始汉语模型层次架构重构的方法来加快维语声学模型的训练速度，还根据最终识别率的高低制定了汉语声学模型层次选择的评估方式，有效地解决了上述三个难点。

3 发明要点

本发明提出一种基于汉语声学模型的维语语音识别系统，来解决维语语音训练数据短缺的问题，并且通过对汉语声学模型层次重构的方法解决了维语声学模型重训练过程中训练速度过慢，训练时间过长的问题，最后，通过一种评估方法解决了基础汉语声学模型层次选择随意性的问题。具体而言，该发明包含如下主要内容：

（1）基于汉语声学模型的维语声学模型自适应方法。

语音识别系统中声学模型的建立需要通过大量的训练来拟合语音信号特征的连续概率分布，无论是传统的混合高斯模型(Gaussian Mixture Model, GMM)，还是现在流行的深度神经网络(Deep neural network, DNN)都需要充分的语料用于训练，否则稀疏的数据无法准确表征语音信号的特征，造成声学模型失配。本发明提出利用 DNN 多层次的特点，将训练充分的汉语 DNN 模型作为基础模型，以此训练维语 DNN 模型，使数据稀疏的维语声学模型自适应，有效降低了维语声学模型的失配度。

（2）基于汉语声学模型的层次重构方法。

在解决维语训练数据稀疏问题时，提到使用汉语 DNN 模型作为基础来解决维语声学模型失配，但汉语 DNN 模型作为基础如何使用，也是一大难点。若将整个汉语和维语 DNN 模型都结合进来，则会造成 DNN 模型结构过于庞大，训练时间过长，训练速度过慢；若只结合部分汉语和维语 DNN 模型，则可能无法实现完全信息共享或者模型结构太过稳定，从而导致系统识别率提升不大。针对这一问题，文发明提出基于汉语声学模型的层次重构方法，将声学模型自适应方法中的模型层次简单组合变成层次重构，可以有效提高语音信息共享，提高维语语音识别系统的识别率。

（3）基于字错误率的层次选择方法。

由于基于汉语声学模型的层次重构方法需要选择汉语及维语 DNN 模型的层次，以组成新的维语声学模型训练架构，因此，层次选择也是一大棘手问题。传统基于大词汇量连续语音识别声学模型重构稀疏训练语料语言声学模型的方法，没有给出选择的评量标准，一般默认选择与基础模型结构一致的层数。我们在层次选择过程中引入字错误率评估方法，允许维语声学模型训练架构变化，即重构的 DNN 模型层数可增、可减，亦可保持不变，而根据各种模型构建的语音识别系统字错误率来选择 DNN 层次结构，从而解决模型层次结构构造随机性问题。

4 发明内容和系统实现

4.1 系统架构

图 1 为维语语音识别系统的结构示意图。该系统分为训练和识别两个子系统。首先需要输入维语训练语音信号，传给训练子系统进行声学模型训练建立维语声

学模型，同时，维语文本输入作为训练子系统中语言模型训练的数据，用以建立维语语言模型；在解码阶段，待识别维语语音信号输入经过训练出的维语声学模型得出大致识别结果，然后结合训练子系统中的维语语言模型得出准确的识别结果。本发明的主要目的是提高在维语训练数据稀疏的条件下，基于 DNN 的维语声学模型表征实际发音准确度，以提升整个维语语音识别系统的识别准确率，因此集中于维语语音识别系统中的维语声学模型训练模块，如图 1 中虚线框所示。

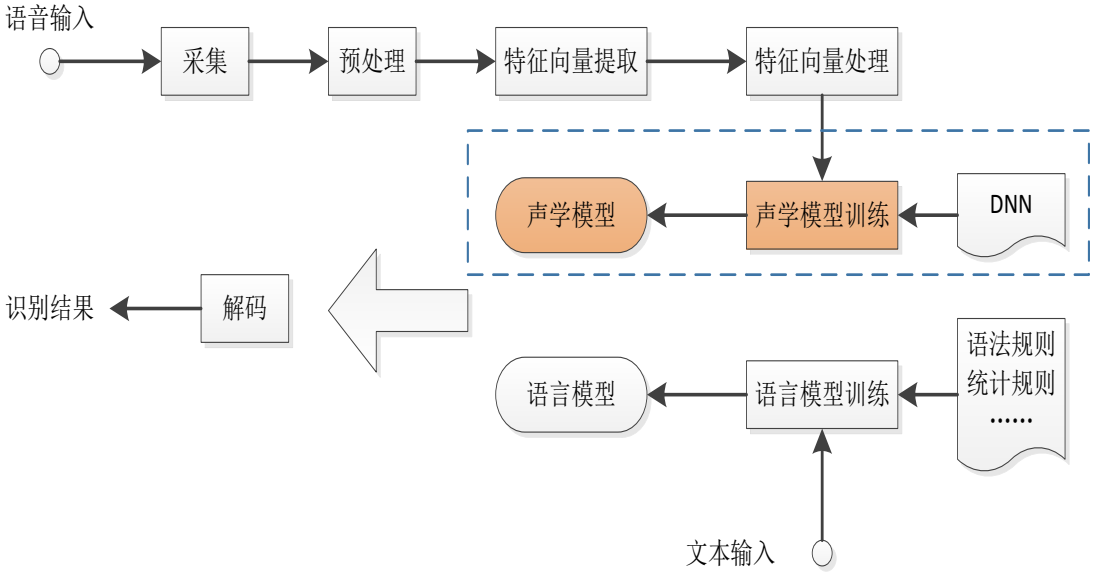

图 1： 维语语音识别系统架构

4.2 基于汉语声学模型的维语声学模型自适应方法

由于深度神经网络(DNNs)的多隐藏层结构，使得 DNN 具有从原始数据中学习层次特征的能力，因此 DNN 在表征语音信号复杂模式上具有高度灵活性，这使得 DNN-HMM 模型代替传统 GMM-HMM 模型在语音识别系统声学模型建模中取得了巨大的成功。图 2 为 GMM 模型与 DNN 模型，图 3 为 DNN-HMM 模型框架图。

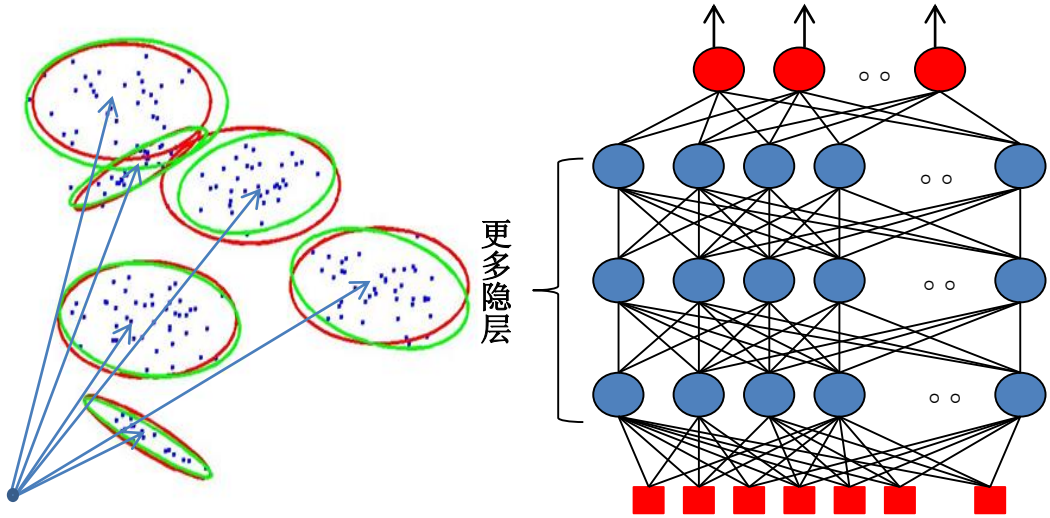

图 2：GMM 模型及 DNN 模型

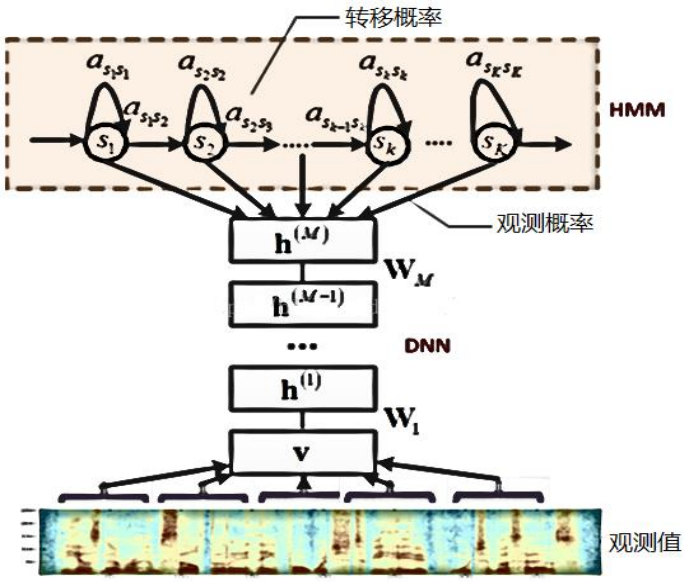

图 3：DNN-HMM 模型框架图

由图 2 可以看出，相比较于传统的 GMM 模型只有均值、方差两个参数，DNN 模型具有多隐藏层（通常大于 3 层）的网络结构，并且每一层都有许多节点（通常上千个），相邻层次间是通过权值 W 全连接的，并且每个隐藏层的输出都经过非线性映射后作为下一层的输入；结合图 3 可发现 HMM 模型中的发射概率是通过 DNN 模型来决定的，而 HMM 模型的转移概率为 DNN 模型提供原型，即音素与共享状态的对应。

从上述分析中可以看出， DNN 模型的本质思想是堆叠多个神经元层，每个层都提取一定的特征和信息，以此来模仿人脑的机制来解释数据，并且，训练过程中，只需指定网络的层数，而不需要给定具体的参数，网络通过计算数据来自动学习最终的参数，不一样的网络参数能够识别不同的物体，训练好的网络就能自动识别物体。因此当数据量过于稀疏时，网络无法学习出语音信号中各种复杂的特征，而导致训练所得声学模型失配。

本发明提出用基于汉语的声学 DNN 模型来自适应维语声学 DNN 模型。一个 DNN 模型可以简单表示为一个三元组 (X, W, K, Y)，其中 $X=(x_1, x_2, x_3 \cdots x_n)$ 为输入向量集合，$W=(w_{j1}, w_{j2}, w_{j3} \cdots w_{jn})$ 为连接权值集合，$K=(1, 2, 3 \cdots k)$ 为隐藏层数，$Y=(y_1, y_2, y_3 \cdots y_m)$ 为输出向量集合。因此一个稳定的声学 DNN 模型，即通过大量 X 训练 W。图 4 所示为训练充分 DNN 抽象识别过程。

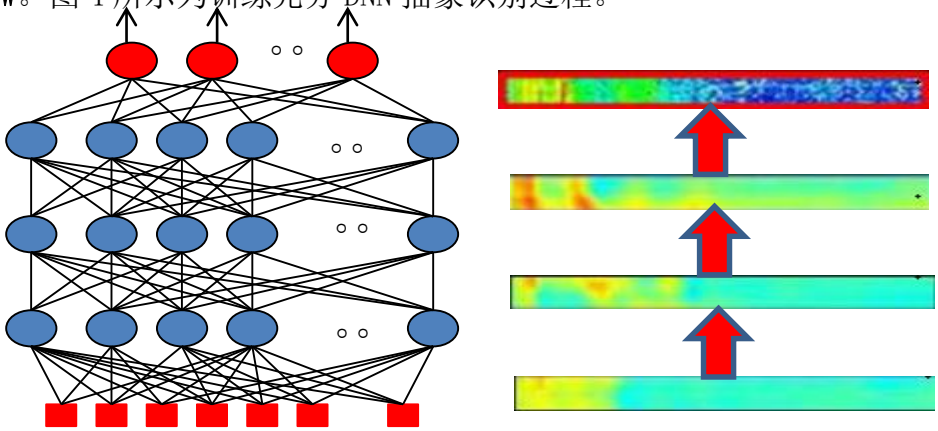

图 4：训练充分 DNN 模型抽象识别过程

通过 DNN，我们可以把多语言的 DNN 模型结合起来，形成一个底部训练充分，抽象特征可供顶部共享的自适应模型，从而解决训练数据过于稀疏导致训练不充分，模型失配的问题。简单而又鲁棒。如图 5，即是通过汉语声学模型构建的自适应维语声学模型。

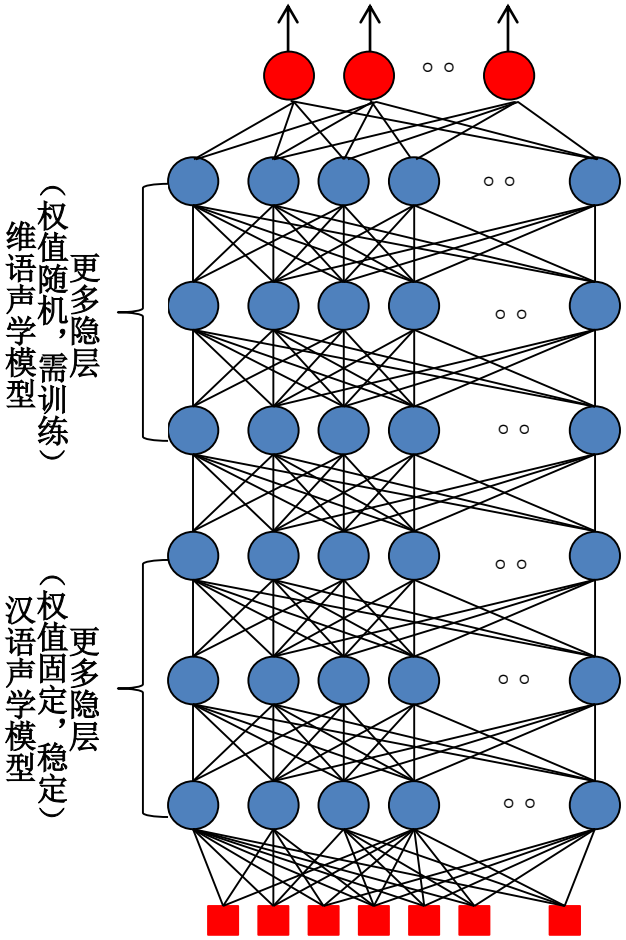

图 5：<汉语>+<维语>自适应声学模型

通过图 5，我们看到由汉语和维语构建的自适应声学模型，底部由汉语抽象出的语音特征都被共享了，这给需训练的维语模型提供了基础和初始依据，能使得新模型在少量维语训练数据的条件下，依然能足够稳定。当维语训练数据少到稀疏时，这种基于汉语声学模型的自适应维语声学模型构造方法将大大优于传统声学模型构造方法。

值得注意的是，基于汉语声学模型的自适应维语声学模型构建方法虽然解决了由于训练数据稀疏导致的声学模型失配问题，但是也带来了新构建的声学模型结构过于庞大的麻烦。在 4.3 节我们将介绍维语自适应声学模型层次重构的方法来解决这一灾难性问题。

4.3 基于汉语声学模型的层次重构方法

基于汉语声学模型的维语声学模型自适应方法可以有效提高维语语音识别系统的识别率，但对于庞大的模型结构却显得无能为力。例如，针对出现的维语方言的二次开发，这些方言的出现需要借鉴现有维语语音识别系统的声学模型再建立自适应模型，这会使得 DNN 声学模型的网络结构继续增长，给训练系统带

来巨大的压力，甚至当维语方言种类增长到一定程度时，系统将完全失效。为了解决这一问题，本发明提出基于汉语声学模型的层次重构方法。自适应模型的层次结构虽然可以无限增多，但基于汉语的共享信息是封闭的，这意味着只要保证用于共享的信息覆盖足够全面，其他稀疏语言都有很大概率以汉语信息为基础构建稳定的声学自适应模型。基于汉语声学模型的层次重构方法不仅可以摆脱自适应模型过于庞大的困扰，还能实现语音信息的全面共享。如图 6 所示，将图 5 中的自适应声学模型按不同方法重构后，保证自适应声学模型层次在正常范围内的同时也保证了汉语声学模型的信息共享，从而保证了系统性能，提高了系统效率。

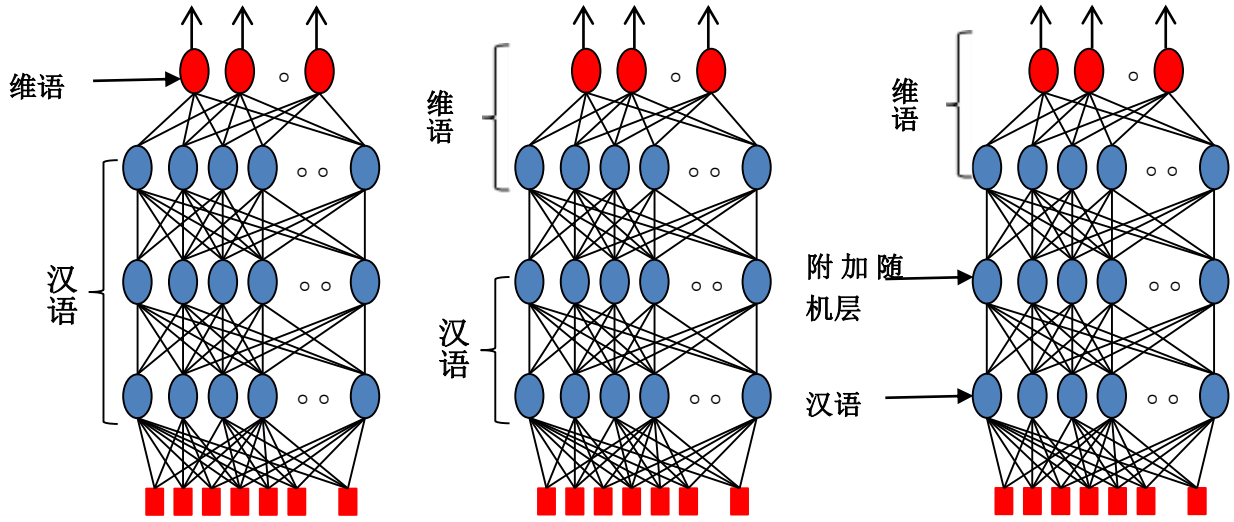

图 6：重构后的自适应声学模型

4.4 基于字错误率的层次选择方法

在 4.3 节中，我们提到可以利用基于汉语声学模型的层次重构方法来优化构建的自适应维语声学模型的层次结构，提高系统的性能和效率。但是重构过程中，如何选择用于重构的部件却没有统一的方法。若选择汉语声学模型的层次过多，维语声学模型的层次较少，则可能导致自适应的维语声学模型训练不够充分，DNN 的灵活性得不到完全施展，这种方式虽然能提高模型的训练效率，但却会使系统的识别性能下降；若选择汉语声学模型的层次过少，维语声学模型的层次较多，则可能导致汉语语音信息不能被自适应维语声学模型完全共享，稀疏的维语训练数据依然不足以训练出一个稳定，覆盖全面的维语声学模型，这种方法也无法使系统识别性能达到令人满意的水准；若选择汉语声学模型层次和维语声学模型层次一样多，则可能导致自适应维语声学模型只能共享部分汉语声学信息，模型也只能部分稳定，这种折衷方法可能使得模型的训练效率和系统识别性能较为平均，但是却无法达到最优目的。

本发明提出基于字错误率的层次选择方法，在层次重构的过程中，重构的自适应模型层次不固定，可增、可减，亦可保持不变，并且能加入随机初始化的随机层，但需保证重构后模型具有较高的训练效率，系统识别性能达到最优。在此过程中，我们引入了一个基于字错误率的评估函数：

$$wer = f(\{p_m\},\{q_n\},\{r_k\}) \qquad p_m, q_n, r_k \in \{0, \qquad (1)$$

其中， $p_m, q_n, r_k \in \{0,1\}$ 分别表示对应的汉语层、随机层、维语层未选中或选中；m, n, k 分别表示可供选择的汉语、随机和维语层次总数。选择过程可分为以下两种情况：

4.3.1 穷举法层次选择

穷举法的基本思想是根据题目的部分条件确定答案的大致范围，并在此范围内对所有可能的情况逐一验证，直到全部情况验证完毕。若某个情况验证符合题目的全部条件，则为本问题的一个解；若全部情况验证后都不符合题目的全部条件，则本题无解。

当原始 DNN 声学模型层次较少（小于 5）时，可简单的选用穷举法。选择过程中，为保证模型训练效率，需保证重构的自适应维语声学模型中隐藏层层数不大于原始模型层次数。因此，只需构建出所有满足条件的自适应维语模型，然后计算每个模型在其他测试条件完全一致的情况下的字错误率，选择字错误率最低的模型即可。

穷举法这种简单的处理过程能给系统设计带来极大的便利，但是随着决定维语语音系统识别性能和模型训练效率的原始 DNN 模型层次的增多，在该系统中时间复杂度趋于 $O(n^3)$的穷举法显然会失效。为此，在处理含有大规模隐藏层的原始模型时，我们选用遗传算法来进行层次选择。

4.3.2 遗传算法层次选择

遗传算法（Genetic Algorithms,GA）是一类借鉴生物界自然选择和自然遗传机制的随机化搜索算法。它模拟自然选择和自然遗传过程中发生的繁殖、交叉和基因突变现象，在每次迭代中都保留一组候选解，并按适度值评估函数从解群中选取较优的个体，利用遗传算子（选择、交叉和变异）对这些个体进行组合，产生新一代的候选解群，重复此过程，直到满足某种收敛指标为止，流程图见图 7。

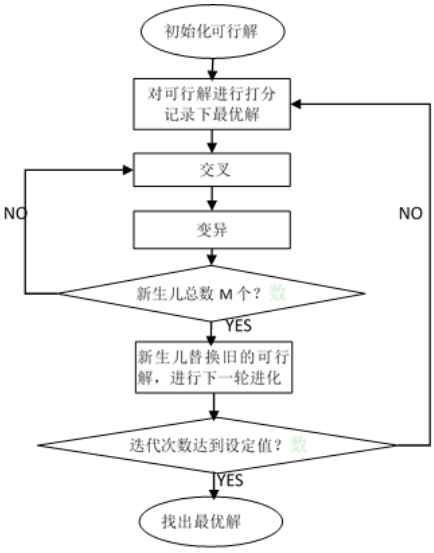

图 7：遗传算法流程图

4.3.1 节中，我们提到，当原始供选择的 DNN 声学模型层次在大规模的条件下时，简单的穷举法无法满足算法的要求，因此，我们需要更加智能的寻优算法来解决层次选择的难题。基本思路是，将选中的层次置为 1，未选中的层次置为 0，满足重构的自适应模型层次总数不大于原始模型层次数，这样每一个选中/未选中序列为遗传算法的一个个体。通过对初始个体附加遗传算子，可实现对个体的“遗传与变异”，可以得到更多个体。对这些个体对应的评估函数值进行比对，直到得到最优个体不在改变，即得到一个最好的变换。

具体方法过程如下：

(1)选中/未选中层次编码。

对应选中/未选中汉语层第 i 层，随机层第 j 层，维语层第 t 层，相应将 p_i、q_j、r_t 置为 1/0，i<=m; j<=n; t<=k，且 0<i+j+t<=原始模型层次数（若三个原始模型层次数不同，则取三者中最大者）。

(2)选择评估函数。

评估函数我们依然选择式(1)所示基于字错误率的评估函数。

(3)算子选择

遗传算法的算子包括选择算子、交叉算子和变异算子。但是在层次选中/未选中编码时，只是对对应的层次置 1/0，而不能随便改变其值。所以在进化过程中，不能使用外变异算子。

1. 选择算子

为了保证算法的全局搜索能力，采用最优个体保存算子，即父代群体中的最优个体直接进入子代群体中，保证遗传过程中所得到的个体不会被交叉和变异操作所破坏。

2. 交叉算子

交叉算子是产生新个体的主要方法，决定了遗传算法的全局搜索能力，在遗传算法中起关键作用。由于层次选中/未选中编码是包括二进制编码变化形式，变化形式比较单一，所以选择简单有效的单点交换算子。

3. 变异算子

变异算子是产生新个体的辅助方法，它决定了遗传算法的局部搜索能力。变异算子和遗传算子相互配合，可以共同完成对搜索空间的全局搜索和局部搜索。为了快速的进行所有满足条件的二进制编码变化，在这里也引入简单的变异算子。

为进一步说明各遗传算子作用下遗传算法的进化过程，我们以汉语、随机、维语各 DNN 模型均为 4 层结构时为例，因此，最终子代选择时，1 的个数大于 4 的编码串将被淘汰。具体过程见图 8。

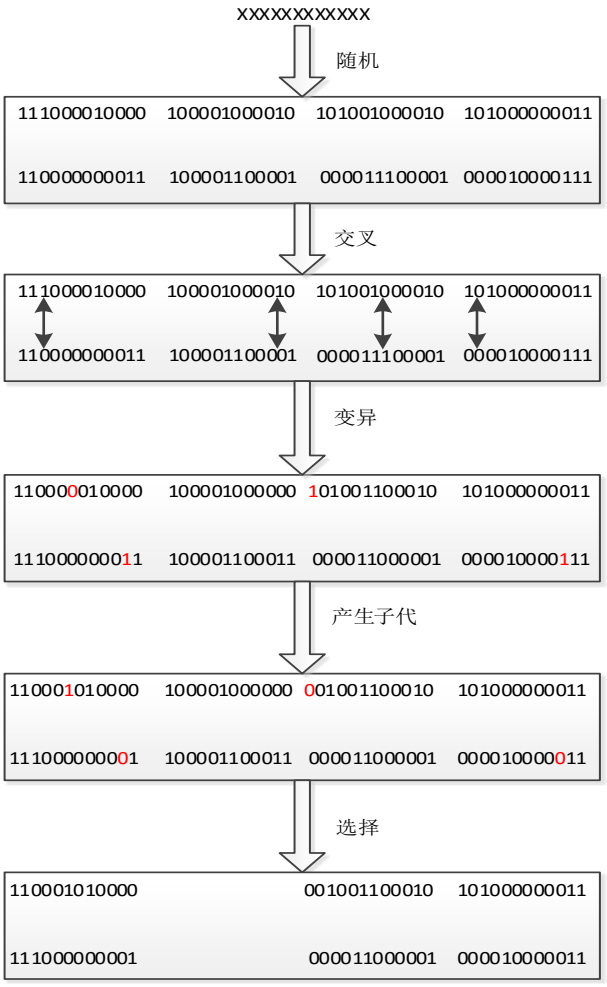

图 8：不同遗传算子的子代产生过程

使用遗传算法来解决大规模的搜索问题，以此求解出问题的最优解，能在很大程度上缓解由于模型规模的增长所带来的时间开销，能有效保证在大规模问题上，该系统依然能保持高的训练效率和识别性能。并且随着原始 DNN 模型层次的增多，算法在该系统上的优越性越能显著的突显出来。

5 发明优势

- 本发明利用利用基于汉语声学模型的维语声学模型自适应方法，极大的解决了由于维语训练数据稀疏而导致的训练所得维语声学模型与实际发音失配的问题，摆脱了以往维语语音识别系统设计中，训练声学模型对数据依赖性过高的问题，减少了数据采集带来的困扰。
- 通过将维语自适应声学模型重构，解决了在基于汉语声学模型的维语声学模型自适应方法中，无法保证自适应模型结构精简，训练效率高，系统识别性能强等问题。
- 基于字错误率的层次选择方法，不仅使得基于汉语声学模型的维语自适应声学模型结构精简，而且还能保证整个维语语音识别系统的高效率与高性能，彻底解决了在稀疏数据下快速训练识别性能高的维语语音识别系统问题。